

Anhedonia-Related Wording in Social Media: An Application of Natural Language Processing

Martin C, Vir M



INTRODUCTION

- Atria identified and collected posts from various mental health communities or 'subreddits' on Reddit.
- Atria aimed to identify a user's mental disorder, focusing particularly on anhedonia, alongside depression, anxiety, bipolar disorder, and borderline personality disorder (BPD).

OBJECTIVES

- Anhedonia is the loss of pleasure or interest in activities that were once enjoyable.
- Traditional diagnostic methods can be subjective and may not accurately capture the complexity of depression.
- To address this need, Atria explored the use of NLP techniques to develop a proof-of-concept project for classifying anhedonia from other types of depression.

METHODS

- All posts within the Reddit anhedonia forum between 2017 and 2022 were downloaded. Existing code was adapted to prepare the data for NLP through tokenization, stop-word removal, and lemmatization.
- Six unsupervised machine learning (ML) algorithms (including sentiment analyses, clustering algorithms, and topic models) were used to identify patterns associated with anhedonia among 9,887 posts, which were then interpreted by a team of humans.

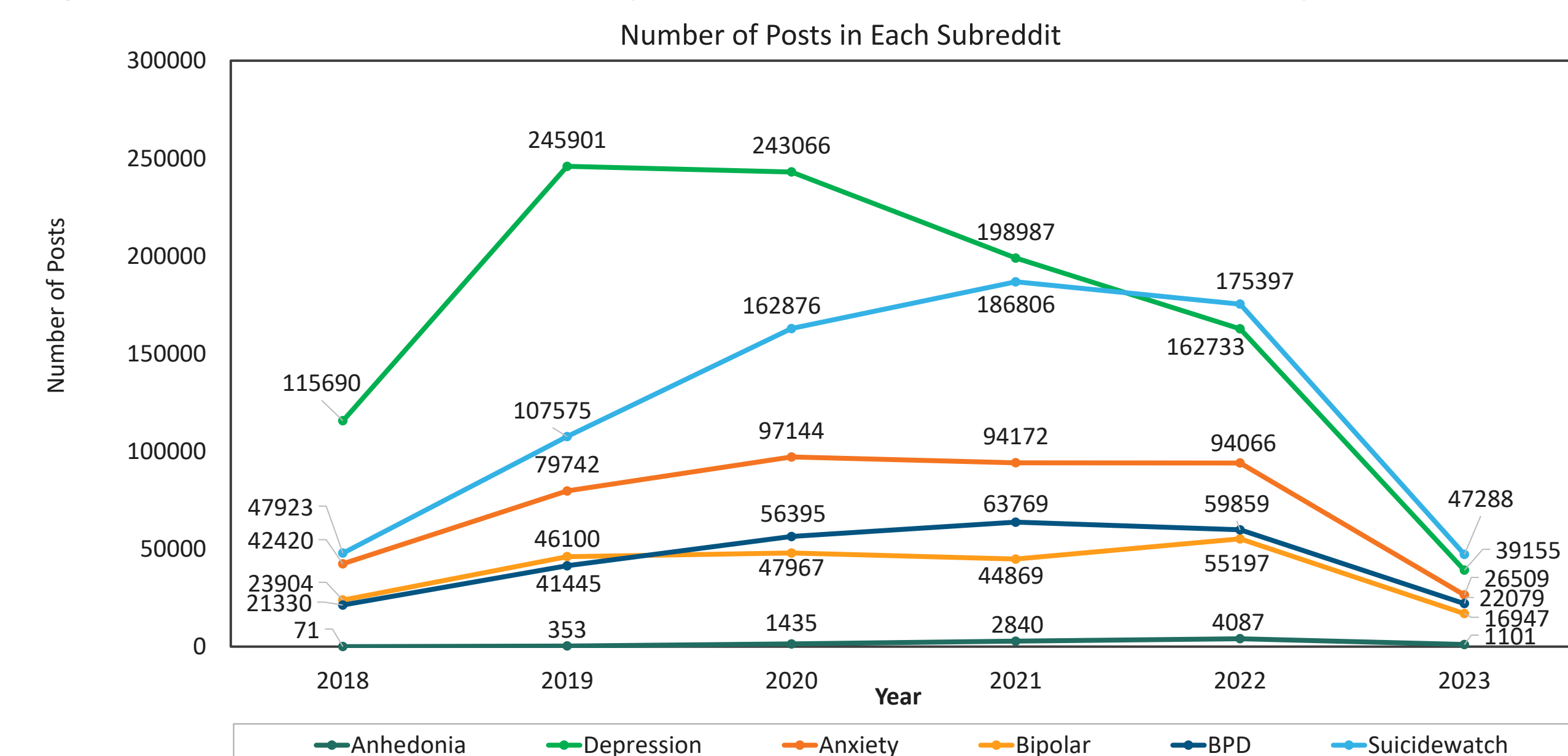
PRELIMINARY ANALYSES

- Atria's preliminary analysis used data from the Depression subreddit between May 2018 and April 2023
- Results showed feasibility of select phrases to capture anhedonia patterns (**Table 1**)
- The number of posts from multiple depression-related subreddits peaked in 2019 and 2020, and has been declining since

Table 1. Summary of Data for Past 5 Years in Each Depression-related Subreddit

Phrases used in depression subgroups (subreddits)*	(N=969,357)
Anhedonia	7,137 (0.7%)
No Interest/ No Motivation/ Lost Interest/ Lack of Interest	12,595 (1.3%)
Fatigue	4,616 (0.5%)
Any of the above	24,348 (2.4%)

Figure 1. Number of Posts in Each Depression-related Subreddit for the Past 5 years



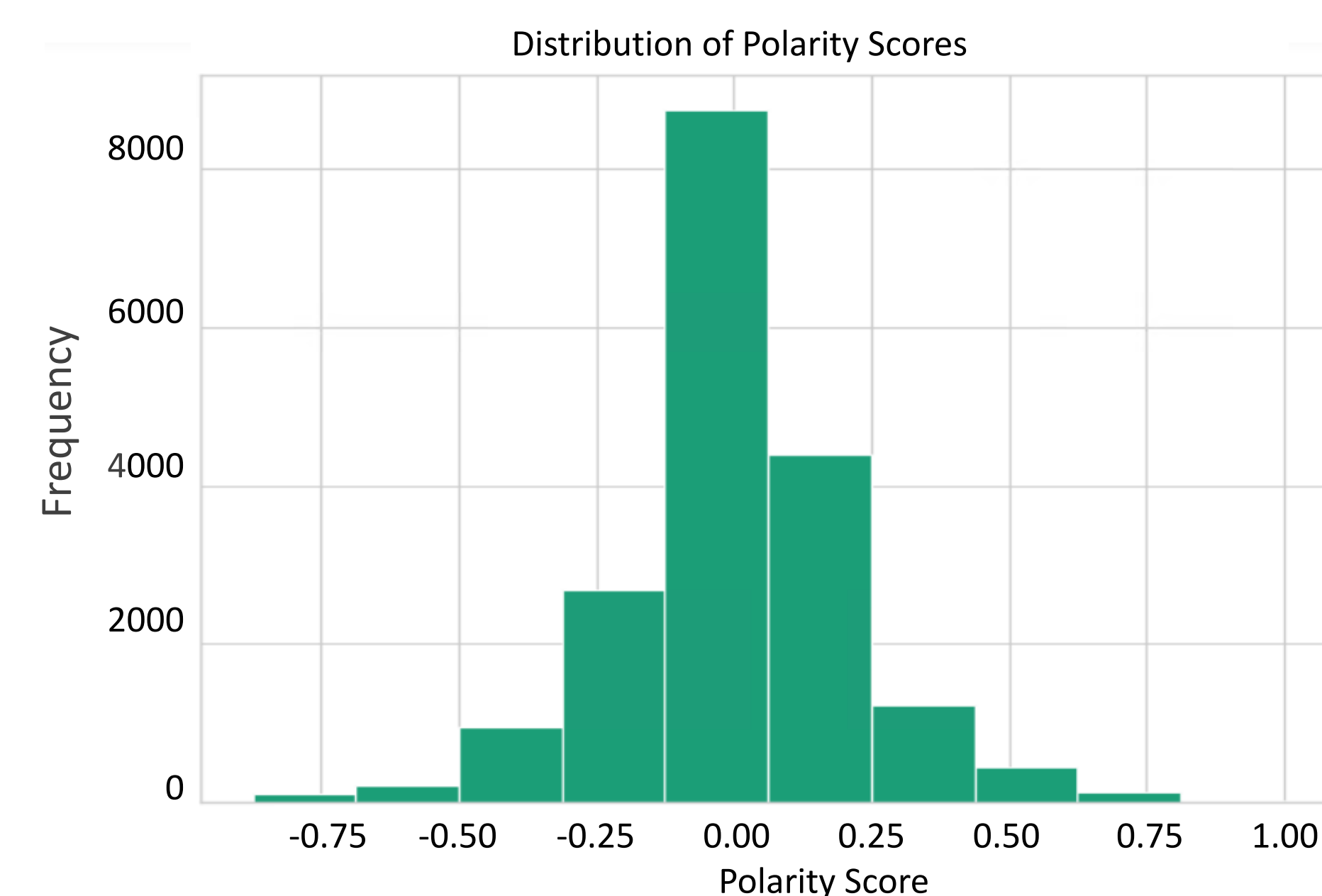
SENTIMENT ANALYSIS

- Textblob** is a python library for performing NLP tasks using a pattern-based approach to determine the sentiment of a piece of text; the presence of certain words and phrases indicates a Positive, Negative, or Neutral sentiment.

- Pros:
 - Simple to use
 - Pre-trained sentiment analyzer
 - Provides a polarity score (from -1 to 1)
- Cons:
 - May not be as accurate
 - Limited ability to capture nuanced language and context

- Figure 2** shows the sentiment for Anhedonia on Reddit is **Neutral**

Figure 2. Sentiment Analysis with TextBlob

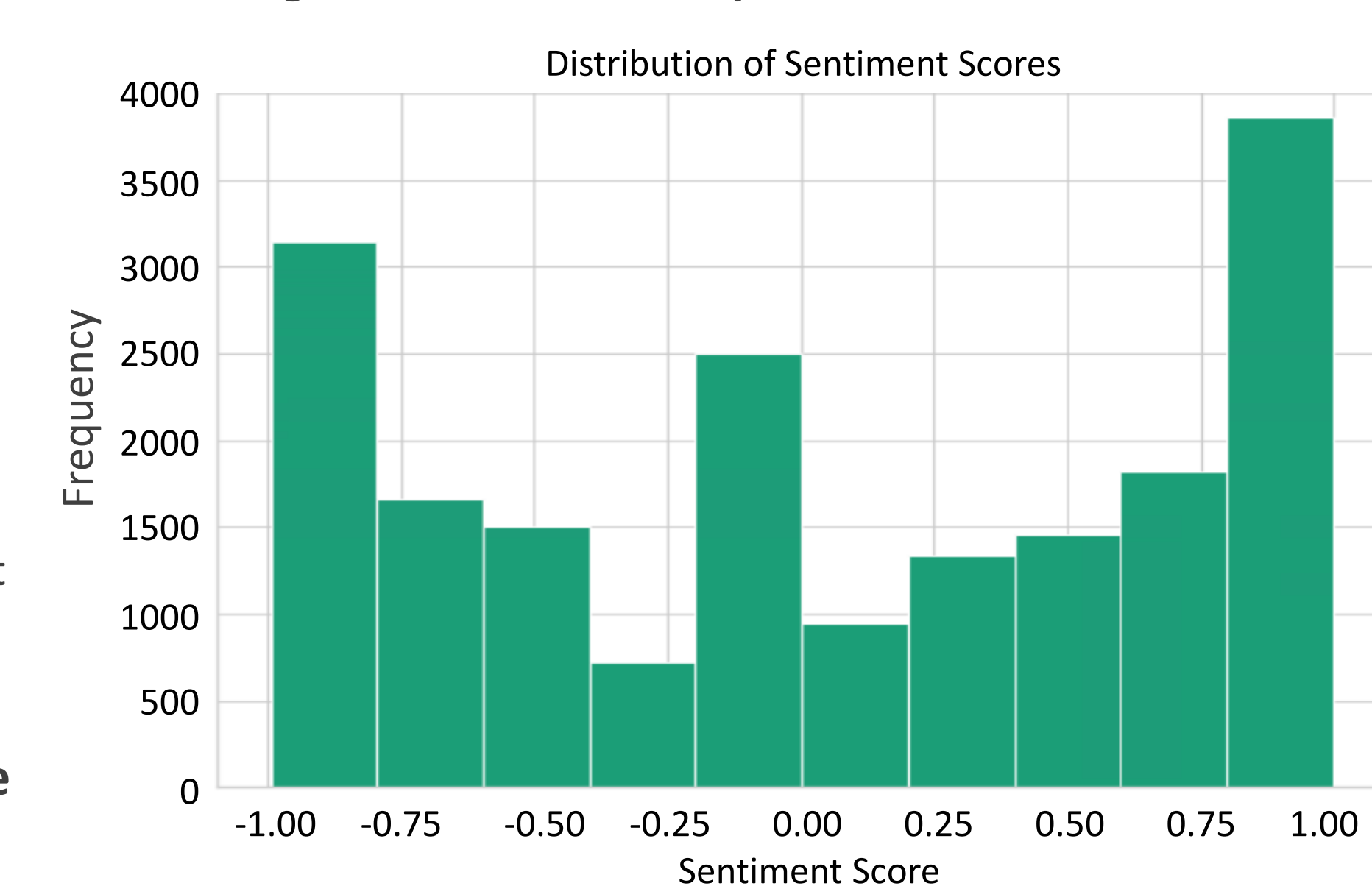


- VADER** uses a combined approach: a pre-defined lexicon of words and phrases, as well as rules that consider context and grammar.

- Pros:
 - Good at capturing sentiment in informal text
 - Performs well in identifying sarcasm and irony
 - Provides a polarity score, as well as separate positive, negative, and neutral scores.
- Cons:
 - May not perform as well on domain-specific text or text outside its pre-defined lexicon

- Figure 3** shows a neutral overall sentiment; however, **positive sentiments are not that positive** while **negative sentiments are very negative**.

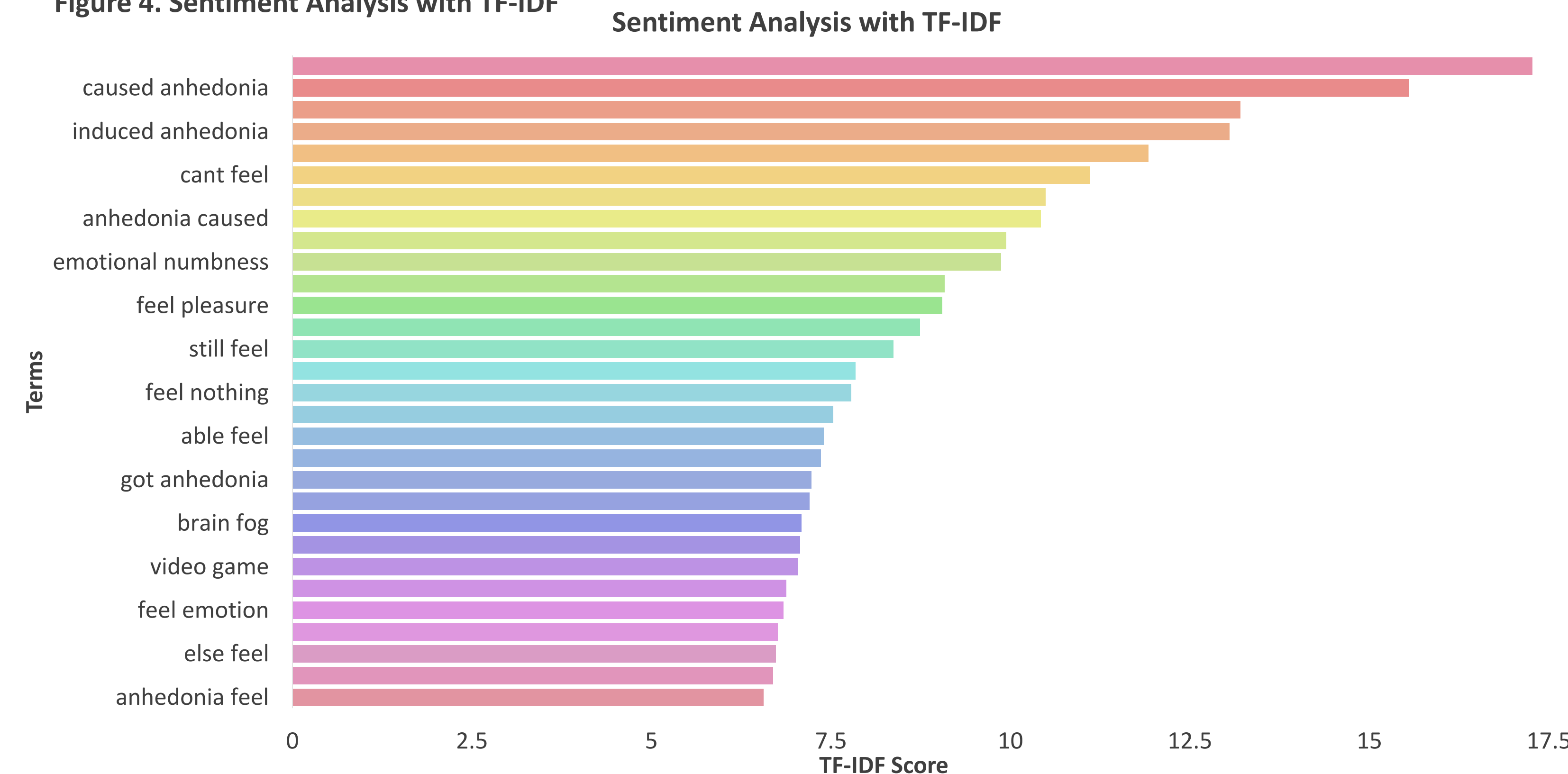
Figure 3. Sentiment Analysis with VADER



TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

- X-axis: TF-IDF (Term Frequency-Inverse Document Frequency) score
 - Relative measure of the importance or relevance of each phrase within the analyzed collection of documents
 - Higher TF-IDF scores indicate a phrase is relatively more significant and distinct
- Broader search using TF-IDF showed "feeling" and "causation" ranked among the most common phrase groupings in posts about anhedonia

Figure 4. Sentiment Analysis with TF-IDF



CLUSTERING

- Clustering** groups similar data points into categories based on similarity. A clustering algorithm might identify a group of posts that discuss feelings of emptiness or lack of motivation, which could be indicative of anhedonia.
- K-means** is a simple and efficient algorithm that can group data points into "k" clusters, where k is the number of clusters specified by the user.
- Initial centroids are chosen randomly from the data points. Each data point is assigned to the nearest centroid based on similarity. Centroids are recalculated until **convergence**. For k=3, we have the following results:

Figure 5. Results for Clustering

Feel Life Enjoy Nothing Day	Anyone Emotion Depress Anhedonia Help	Think People Anything Like Brain
---	--	--

TOPIC MODELING

Latent Dirichlet Allocation (LDA)

- A generative probabilistic model that aims to uncover the underlying topics in a corpus of text
- Helps with the assignment of documents to those topics based on the distribution of words in the document

Non-Negative Matrix Factorization (NMF)

- Technique that aims to factorize the term-document matrix into a lower-dimensional representation of topics, and a set of weights, that represent the importance of each topic in each document

Figure 6. Topic Modeling with LDA

Feel Like Life Emotion Enjoy	Help Anhedonia Dopamine Low Die
People Relation Job Like Please	Year Meme Week Know Day

Figure 7. Topic Modeling with NMF

Day Thing Time Year Month	Feel Like Bodies Know Anyone
Anhedonia Depress Dopamine Pleasure Low	Emotion Memory Lost Remember Anymore

CONCLUSIONS

- Applying NLP-based methods to social media revealed anhedonia-associated language. Words strongly linked to the topic of anhedonia could help identify anhedonia-related content in other datasets.
- Future analyses may help identify undiagnosed individuals using informed models and large standardized datasets such as electronic health records and patient health questionnaires.
- For next steps, Atria seeks to:
 - Improve pre-processing
 - Increase data size from 20,000 posts to one million posts
 - Improve feature extraction
 - Implement neural network and supervised learning to classify anhedonia from other types of depression

REFERENCES

- Devika, M.D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science, 87, 44-49. <https://doi.org/10.1016/j.procs.2016.05.124>
- Kim, S.W., & Gil, J.M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences, 9(1), 30. <https://doi.org/10.1186/s13673-019-0192-7>
- Asmussen, C.B., Möller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data 6, 93 (2019). <https://doi.org/10.1186/s40537-019-0255-7>
- K. Bindra and A. Mishra, "A detailed study of clustering algorithms," 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2017, pp. 371-376, doi: 10.1109/ICRITO.2017.8342454.

DISCLOSURES

- CM and VM are employees of Atria.
- This study was fully funded by Atria.

POSTER PRESENTED AT

ISPOR US, Atlanta, Georgia, USA; 5 May – 8 May 2024.

